(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 January 2001 (18.01.2001)

PCT

(10) International Publication Number
WO 01/05107 A1

(51) **International Patent Classification**⁷: **H04L 12/56**

(21) **International Application Number:** PCT/US00/19006

(22) **International Filing Date:** 13 July 2000 (13.07.2000)

(25) **Filing Language:** English

(26) **Publication Language:** English

(30) **Priority Data:**
60/143,431 13 July 1999 (13.07.1999) US

(71) **Applicant** *(for all designated States except US)*: **ALTEON WEB SYSTEMS, INC.** [US/US]; 50 Great Oaks Road, San Jose, CA 95119 (US).

(72) **Inventor; and**
(75) **Inventor/Applicant** *(for US only)*: **KONG, Cheng-Gang** [US/US]; 19591 Moray Court, Saratoga, CA 95070 (US).

(74) **Agents: GLENN, Michael, A.** et al.; Glenn Patent Group, 3475 Edison Way, Ste. L, Menlo Park, CA 94025 (US).

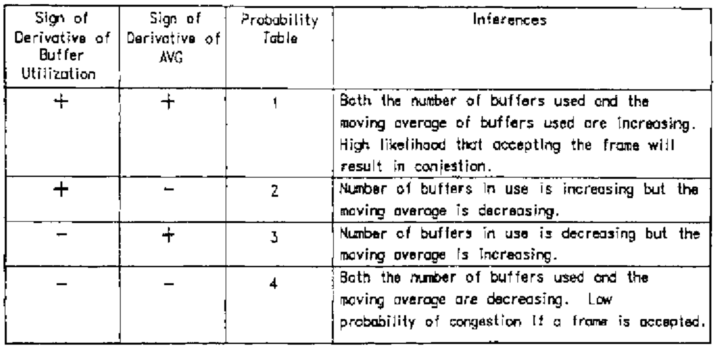
(81) **Designated States** *(national)*: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW.

(84) **Designated States** *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *With international search report.*
— *Before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments.*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) **Title:** APPARATUS AND METHOD TO MINIMIZE CONGESTION IN AN OUTPUT QUEUING SWITCH



Sign of Derivative of Buffer Utilization	Sign of Derivative of AVG	Probability Table	Inferences
+	+	1	Both the number of buffers used and the moving average of buffers used are increasing. High likelihood that accepting the frame will result in conjestion.
+	−	2	Number of buffers in use is increasing but the moving average is decreasing.
−	+	3	Number of buffers in use is decreasing but the moving average is Increasing.
−	−	4	Both the number of buffers used and the moving average are decreasing. Low probability of congestion If a frame is accepted.

(57) **Abstract:** Computer network switching units have limited hardware resources. When a switching unit can not accommodate the aggregate data load arriving at its input, it must drop data frames that it can not forward. Switching units will normally set two thresholds that are compared to the current state of hardware utilization in the switching unit. The first threshold indicates that the utilization of hardware is low. If current utilization falls below this first threshold, the inference is that the switching unit can pass a data frame. If current hardware utilization exceeds an upper threshold. then the inference is that the switching unit is saturated and can not effectively pass a data frame. When current hardware utilization is found to be between these two threshold limits, the switching unit relies on one of four probability tables to decide if the data frame should be dropped. Thè values in these tables are established empirically. The probability values in these four tables define the probability that a data frame will be lost given the current level of hardware utilization. The switching unit maintains historical trends of utilization and uses the sign of the first derivative of the historical data and the sign of the first derivative of a filtered version of the historical data to predict if hardware utilization is increasing or decreasing. Using the signs of these two first derivatives to select one of the four tables, the switching unit can then read a probability value from one of the tables. If that value exceeds a third threshold, the switching unit decides to drop the data frame.

WO 01/05107 A1

APPARATUS AND METHOD TO MINIMIZE CONGESTION IN AN OUTPUT QUEUING SWITCH

BACKGROUND OF THE INVENTION

TECHNICAL FIELD

The present invention relates to the processing and management of data flowing through a computer network switch.

DESCRIPTION OF THE PRIOR ART

Computer networks are constructed by tying together a plurality of switching units. The switching units receive data from various sources in a quantum known as a frame. As computer networks continue to proliferate throughout the world, switching units must be able to route an ever-increasing bandwidth of data. As the bandwidth increases, switching units must be able to handle a greater number of data frames per given unit of time.

The switching units themselves rely on various strategies to ensure that the total frame rate can be accommodated. In the previous art, as frame rate increased the hardware foundation of the switching unit would be taxed to such an extent that some frames would inevitably be lost. These lost frames are known as dropped frames.

Switching units don't just drop frames. Frames are dropped intentionally and systematically based on a set of criteria that reflects the current utilization of all of the resources in the switching unit. Some switching units known today wait until the data frames undergo a process known as "forwarding" before the decision to drop a frame is made. This means that the decision point occurs after the frame is queued up for output to a communications channel.

Contention for available resources in the switching unit causes a state of congestion through the switching unit. To reduce the congestion, the switching unit executes a process that monitors the total amount of data traffic currently being handled and creates historical traffic patterns that it uses to predict future contention levels. These techniques are collectively known as active queue management methods.

One popular congestion prediction mechanism employed by prior art switching units is known as the Random Early Drop method. The Random Early Drop method compares current resource demand against two thresholds; high-threshold and low-threshold. The resource that is monitored is utilization of queue output buffers. In order to reduce the noise associated with the bursty use of these output buffers, the utilization rate is first subjected to a low pass filter. It is the output of the low pass filter that is actually compared against the two thresholds.

If the output of the low pass filter is greater than the high-threshold, the data frame is dropped. Conversely, data frames are never dropped if the low-threshold level is not reached. When the filtered buffer utilization rate is found to be between the two thresholds, a table is consulted to determine the probability that the new data frame will cause congestion. The probability tables are indexed by comparing the difference between the low-threshold and the filtered output. The probability tables specify the likelihood of congestion based on a-priori knowledge of the switching unit's capabilities.

SUMMARY OF THE INVENTION

The methods and apparatus described herein implement a novel and unique facility that provides for significant improvement in the actual rate at which data frames are dropped by a computer network switching unit. By generating and continuously updating utilization histograms, the switching unit can anticipate utilization of switching unit resources. The switching unit uses a new congestion controller that considers the first derivatives of the real-time utilization and of a filtered rendition of the utilization histograms. The sign of these two derivatives define four states of utilization wherein each state carries an inference of upcoming changes in utilization. The new congestion controller uses these four states to select one of four probability tables. The congestion controller reads a probability value from one of these tables to determine if a data frame should be dropped.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a function flow diagram that depicts the processing performed by the Prior Art Random Early Drop method of congestion predication;

Fig. 2 is a graph that presents a histogram for output queue buffer volume;

Fig. 3 is a table that defines the four states that the variance in queue buffer utilization and the average thereof can assume;

Fig. 4 is a block diagram that depicts the preferred hardware embodiment of the present invention;

Fig. 5 is a block diagram that depicts the preferred hardware embodiment of the congestion controller integral to the present invention;

Fig. 6A is the first portion of a Flow diagram that depicts the sequence of steps that the congestion controller follows when determining if a data frame should be dropped; and

Fig. 6B is the second portion of a Flow diagram that depicts the sequence of steps that the congestion controller follows when determining if a data frame should be dropped.

DETAILED DESCRIPTION OF THE INVENTION

The present invention is a method and an apparatus that minimizes the loss of data flowing through a switching unit based on a novel method of congestion control. The apparatus is embodied as a queue switch.

Prior Art

In order to fully appreciate the utility of the present invention, it is necessary to review the prior art related to congestion control and most specifically the Random Early Drop (RED) method of congestion prediction.

Fig. 1 depicts the processing performed by the RED method. Real-time buffer utilization rates are indicative of the number of output queues that will be required to send output data to a communications channel. As can be seen in this figure, the real time buffer utilization rates 5 are directed to a low pass filter 10. The low pass filter 10 is implemented in software that executes in the control processor of a switching unit. The low pass filter 10 is best implemented as an exponential weighted moving average, the output of which is referred to by the mnemonic AVG. The moving average filter reduces short-term variations in the buffer utilization rate. This provides a more stable statistical basis for the congestion prediction method.

The output of the low pass filter 10 is called the average buffer utilization rate 15 (AVG). In the prior art RED method, the value of AVG 15 is immediately compared to two thresholds; high-threshold and low-threshold. The comparison is made in software, but it is functionally depicted in this figure as two hardware comparators, 20 and 25 respectively. The output of the high-threshold comparator 20 is used to determine if the incoming data frame must be dropped. If the value of AVG 15 is less than the low-threshold, the low threshold comparator 25 indicates that the frame can be accepted without increased contention for the switching unit's hardware resources.

If neither the drop-frame indicator 30 or the accept-frame indicator 35 are active, as detected by an AND gate 40, the value of AVG 15 is summed with the negative of the low-threshold. The difference of these two values, which is called the table index 45, is used to select a probability value from a probability table 50. The probability table 50 is filled with values that are empirically determined by monitoring the performance of the switching unit under varying data loads with the congestion control methods disabled. The output of the probability table 50 is then compared against an acceptable probability threshold 55 in order to make the drop-accept decision.

Improved Method

The prior art is an effective congestion control method, but it has limited value because its prediction is based only on the filtered value of the data frame volume. The two thresholds are an adequate means for establishing a straight frame dropping criteria, but the intermediate values of AVG 15 that lie inside the two thresholds add little to no benefit aside from serving as an index into the probability table 50. The improved method considers the historical trend of resource utilization in order to provide additional table indexing.

Real-time buffer utilization rates can change almost instantaneously, i.e. in step functions. This type of traffic pattern is often referred to as being bursty. The filtered rendition of the real-time buffer utilization rate does not exhibit these step functions, rather it follows a smoother curve commensurate with the filtering function provided by the low pass filter 10.

The prior art used the filtered value of the real-time buffer utilization, AVG 15, alone to index a table of congestion probabilities. The present invention uses not only the AVG 15 value, but also considers the direction of the change in the value of AVG 15 to provide additional frame-drop decision criteria. The present invention also considers the direction of change in the unfiltered real-time buffer utilization.

Fig. 2 presents a typical histogram for network data frame volume. The gray bars 65 record the number of queue buffers being used by the switching unit per unit of time (such as 5,000 per second). The output of the low pass filter is recorded as a curve 70. The present invention exploits the predictive nature of the filter output 70 in that the sign of the first derivative of this AVG curve indicates if the average of the buffer utilization rate is increasing or decreasing. The present invention also determines the sign of the first derivative of the total number of output buffers being used per unit time. This latter characteristic indicates if the utilization of buffers is declining or increasing. Any values specified in this paragraph or that can be inferred from Fig. 2 are for purposes of illustration only. Actual values are dependant on actual load conditions that the switching unit is exposed to.

Fig. 3 presents a table that defines the four states that are defined by the polarity of the first derivative of the quantity of buffers used and the first derivative of the AVG moving average of the quantity of buffers used. In effect, these two first derivatives predict the direction of change in hardware utilization based on the historical trend. A key feature of the present invention is the use of this trend-based prediction to distinguish states of hardware utilization. By distinguishing the state of hardware utilization, the improved method of congestion control can select one of four probability tables instead of one.

In practice, the four probability tables, referenced herein as PT-1 through PT-4, inclusive, are populated with probability values that are discovered through an empirical process. This process involves subjecting the switching unit to varying load conditions that result in the four states defined in Fig. 3. The switching unit is then operated in each of these four states with the congestion control process disabled. The actual drop rate for data frames is recorded for each state at varying levels of the AVG moving average. Basic statistical methods are used to develop drop probability values for each state at the various AVG levels that index the probability table. The probability threshold that is compared against the values stored in the tables can be derived empirically. As an added refinement to this method, the probability threshold can be generated in a random fashion in order to approximate the actual random nature of network loading.

Preferred Embodiment

This improved method of congestion control is best reduced to practice in the form of computer based signal processing. On a periodic basis, a processing element maintains the history of queue utilization. From this history, the processing element calculates a moving average of queue utilization and differentiates both the raw queue utilization function and the filtered moving average function.

Fig. 4 presents the preferred embodiment of the new switching unit including a new congestion controller 100. In operation, Switching unit 95 receives data from an external source through a media attachment unit 105. The media attachment unit 105 generally receives serial data, although the data can be parallel. After receiving data from the external source, the media attachment unit 105 creates data frames that it then presents to a input first-in-first-out (FIFO) buffer 110.

The input FIFO 110 forwards each data frame that it receives to a forwarding engine 115. The forwarding engine 115 determines what output queue each data frame must be directed to in accordance with either *a-priori* routing knowledge or dynamic maps that it creates. Once the forwarding engine 115 has processed a data frame, it is stored in a queue memory 120. The forwarding engine 115 identifies the data frame and the queue to which it was directed and delivers this identification to a queue linker 125. The queue linker 125 informs the congestion controller 100 that a new queue buffer has been allocated. If the congestion controller 100 determines that the data frame should be dropped, queue linker 125 removes the data frame from the processing stream and frees the associated data block in the queue memory 120. Otherwise, the queue linker 125 notifies a switch queue 130 that the queue can be transmitted.

Once the switch queue acknowledges the new queue, it retrieves the data frame from the memory element 120 and delivers it to a switch media attachment unit (MAC) 135.

Figure 5 is a block diagram that depicts the construction of the congestion controller 100, The congestion controller 100 is comprised of a high speed processing element 150, a firmware storage memory 155, a history memory 170 and a probability table memory 180. A regular central processor unit (CPU) or a digital signal processor (DSP) can be used in this application. The CPU, or in the alternate a DSP, executes a series of instructions stored in a firmware storage memory 155.

The congestion controller 100 is further comprised of an input port 160 and an output port 165. The input port is used by the processing element 150 to detect when a buffer has been allocated. A signal is received from the queue linker 125 that indicates when buffers are allocated. This signal is then captured by the input port 160 and conveyed to the processing element 150. After the processing element 150 has determined that a data frame should be dropped, it sends a drop-frame signal to the switch queue 130 using the output port 165.

Fig. 6A and Fig. 6B demonstrate the functional flow of the instruction sequence stored in the firmware storage memory. Once the processing element has sensed the buffer allocation signal (step 200), it begins the process of creating a histogram of buffer allocations (step 205). This is stored in a history memory 170 as a function of time; *B(t)*. The processing element 150, based on the history of the buffer allocation, creates a moving average of the buffer allocation function (step 210). This moving average can be any suitable moving average method. The moving average is referred to by the mnemonic AVG.

On a periodic basis, the period of which is established empirically to maximize the throughput of the switching unit, the processing element 150 executes a series of instructions that effectively differentiates the buffer allocation function stored in the history memory 170 (step 215). The resultant first derivative of the buffer allocation function is also stored in the history memory 170. The processing element 150 then executes a series of instructions that differentiate the moving average of the buffer allocation function (step 220). The resultant first derivative of the moving average of the buffer allocation function is also stored in the history memory 170.

The processing element 150 maintains upper and lower threshold values in a probability table memory 180. These are referred to as T_U and T_L respectively. Whenever the switching unit must decide if a data frame should be dropped, the value of the moving average of buffer allocation (AVG) is compared to the upper and lower thresholds. If the value of the AVG exceeds the upper threshold T_U (step 225), then the processing element uses the output port 165 to signal the switch queue 130 that the data frame should be dropped (step 230). If the value of AVG is less than the lower threshold (step 235), the processing element does not perform any other processing for the current data frame and the data frame is not dropped. This method is analogous to the prior art.

Fig. 6A shows that, in the present art, the processing element 150 performs additional processing to determine if a data frame should be dropped. The processing element uses the sign of the first derivatives of the buffer allocation function and the sign of the first derivative of the moving average to select one of four probability tables stored in probability table memory 180 (step 240). The table selection is made according to the combinations described in Fig. 3. If the value of the AVG is greater than the lower threshold, as determined by inference by step 235, then the processing element 150 then subtracts the value of the lower threshold T_L from the moving average AVG (AVG - T_L) (step 245). The difference of AVG - T_L is used as an index into the selected probability table according to the statement:

$$P_{B(t)';AVG'}(AVG-T_L)$$

where P is a probability table selected by the sign of the first derivative of the buffer utilization function and the sign of the first derivative of the buffer utilization moving average AVG.

Once the processing element 150 has read a probability value from one of the four probability tables stored in the probability table memory 180 (step 250), it then compares that value to an empirically established probability threshold T_p. If the table value exceeds the probability threshold (step 255), then the processing element 150 uses the output port 165 to indicate to the switch queue 130 that the current data frame must be dropped (step 260). In a refinement to the present embodiment, the probability threshold for this comparison can be derived in a random manner to more closely approximate the random nature of actual network loading.

Alternative Embodiments

The key essence of the present invention is the use of the trend analysis mechanism to predict the direction of change in buffer utilization and the moving average thereof. Many alternative embodiments have been considered by the inventor including, but not limited to, using a multi-dimensional table for the storage of empirically discovered probability values. In such an alternative embodiment, the four tables discussed herein are replaced with one table having three indices. A value from such a table would be referenced by the statement:

$$P[B(t)', AVG', (AVG - T_L)]$$

Two tables could be used to store probability values with each table having two indices. An example of such a table reference having two tables selected by the sign of the first derivative of the unfiltered moving average would be:

$$P_{B(t)'}[AVG', (AVG - T_L)]$$

All of the probability tables have been described with an index that represents the difference between the average buffer utilization AVG and the Lower threshold T_L. The index to any of these tables can be the difference of the upper threshold T_H and the average buffer utilization AVG.

CLAIMS

1. A method for reducing resource congestion in an output queuing switch comprising the steps of:

monitoring the utilization of output queue buffers;

calculating a moving average of the utilization of output queue buffers;

accepting a data frame if the moving average is less than a first lower threshold;

dropping a data frame if the moving average is greater than a first upper threshold;

determining the sign of the first derivative of said utilization of output queue buffers;

determining the sign of the first derivative of said moving average;

selecting a probability table based on said sign of the first derivative of said utilization of output queue buffers and said sign of the first derivative of said moving average;

using the difference between said moving average and said lower threshold as an index to read a value in said selected probability table; and

dropping the frame if said value read from said selected probability table according to said index exceeds a pre-determined value and the value of said moving average is greater than said first lower threshold.

2. The method of Claim 1 wherein the pre-determined value is established in a random manner.

3. The method of Claim 1, wherein the first lower threshold in the step of accepting a data frame if the moving average is less than a first lower threshold is established empirically.

4. The method of Claim 1, wherein the first upper threshold in the step of dropping a data frame if the moving average is greater than a first upper threshold is established empirically.

5. The method of Claim 1, wherein the values stored in said probability tables are established empirically.

6. The method of Claim 1, wherein the first lower threshold in the step of accepting a data frame if the moving average is less than a first lower threshold is established analytically.

7. The method of Claim 1, wherein the first upper threshold in the step of dropping a data frame if the moving average is greater than a first upper threshold is established analytically.

8. The method of Claim 1, wherein the values stored in said probability tables are established analytically.

9. A method for reducing resource congestion in an output queuing switch comprising the steps of:

monitoring the utilization of output queue buffers;

calculating a moving average of the utilization of output queue buffers;

accepting a data frame if the moving average is less than a first lower threshold;

dropping a data frame if the moving average is greater than a first upper threshold;

determining the sign of the first derivative of the utilization of output queue buffers;

determining the sign of the first derivative of the moving average;

using the difference between said moving average and said lower threshold together with said sign of the first derivative of said utilization of output queue buffers and said sign of the first derivative of said moving average as indices to read a value in a probability table; and

dropping the frame if said value read from said probability table according to said indices exceeds a pre-determined value and the value of said moving average is greater than said first lower threshold.

10. The method of Claim 9 wherein the pre-determined value is established in a random manner.

11. The method of Claim 9, wherein the first lower threshold in the step of accepting a data frame if the moving average is less than a first lower threshold is established empirically.

12. The method of Claim 9, wherein the first upper threshold in the step of dropping a data frame if the moving average is greater than a first upper threshold is established empirically.

13. The method of Claim 9, wherein the values stored in said probability tables are established empirically.

14. The method of Claim 9, wherein the first lower threshold in the step of accepting a data frame if the moving average is less than a first lower threshold is established analytically.

15. The method of Claim 9, wherein the first upper threshold in the step of dropping a data frame if the moving average is greater than a first upper threshold is established analytically.

16. The method of Claim 9, wherein the values stored in said probability tables are established analytically.

17. An output queuing switch apparatus comprising:

network receiver circuit that accepts data from an external data source;

wire-input first-in-first-out buffer that accepts data from said network receiver circuit and assembles data frames;

memory element;

forwarding engine that receives data frames from said wire input first-in-first-out buffer, determines the appropriate destination output queue and stores data frames in said memory element;

queue linker that receives data frame descriptors from said forwarding engine and creates output queues containing said data in said memory element;

congestion controller that:

monitors the utilization of output queue buffers;

calculates a moving average based on the utilization of output queue buffers;

issues a discard signal if said moving average exceeds a first upper threshold;

determines the sign of the first derivative of the utilization of output queue buffers;

determines the sign of the first derivative of said moving average;

selects a probability table based on the said sign of the first derivative of said utilization of output queue buffers and said sign of the first derivative of said moving average;

issues a discard signal if said moving average is greater than a first lower threshold and the value stored in said selected probability table as indexed by the difference of said moving average and said first lower threshold exceeds a first probability threshold;

switch manager that accepts queue data from said memory and discards data frames in response to a discard signal issued by said congestion controller; and

switch media access controller that accepts queue data from said switch manager and dispatches the queue data to a network port .

18. The apparatus of Claim 17, wherein the first lower threshold used in the congestion controller is established empirically.

19. The apparatus of Claim 17, wherein the first upper threshold used in the congestion controller is established empirically.

20. The apparatus of Claim 17, wherein the values stored in said probability tables used in the congestion controller are established empirically.

21. The apparatus of Claim 17, wherein the first lower threshold used in the congestion controller is established analytically.

22. The apparatus of Claim 17, wherein the first upper threshold used in the congestion controller is established analytically.

23. The apparatus of Claim 17, wherein the values stored in said probability tables used in the congestion controller are established analytically.

24. The apparatus of Claim 17 wherein said first probability threshold is established empirically.

25. The apparatus of Claim 17 wherein said first probability threshold is established in a random manner.

26. An output queuing switch apparatus comprising:

network receiver circuit that accepts data from an external source;

wire-input first-in-first-out buffer that accepts data from said network receiver circuit and assembles data frames;

memory element;

forwarding engine that receives data frames from said wire input first-in-first-out buffer, determines the appropriate destination output queue and stores data frames in said memory element;

queue linker that receives data frame descriptors from said forwarding engine and creates output queues containing said data in said memory element;

congestion controller that:

monitors the utilization of output queue buffers;

calculates a moving average based on the utilization of output queue buffers;

issues a discard signal if said moving average exceeds a first upper threshold;

determines the sign of the first derivative of the utilization of output queue buffers;

determines the sign of the first derivative of said moving average;

issues a discard signal if said moving average is greater than a first lower threshold and the value stored in a probability table as indexed by the difference of said moving average and said first lower threshold together with the said sign of the first derivative of said utilization of output queue buffers and said sign of the first derivative of said moving average exceeds a first probability threshold;

switch manager that accepts queue data from said memory and discards data frames in response to a discard signal issued by said congestion controller; and

switch media access controller that accepts queue data from said switch manager and dispatches the queue data to a network port .

27. The apparatus of Claim 26, wherein the first lower threshold used in the congestion controller is established empirically.

28. The apparatus of Claim 26, wherein the first upper threshold used in the congestion controller is established empirically.

29. The apparatus of Claim 26, wherein the values stored in said probability table used in the congestion controller are established empirically.

30. The apparatus of Claim 26, wherein the first lower threshold used in the congestion controller is established analytically.

31. The apparatus of Claim 26, wherein the first upper threshold used in the congestion controller is established analytically.

32. The apparatus of Claim 26, wherein the values stored in said probability table used in the congestion controller are established analytically.

33. The apparatus of Claim 26 wherein said first probability threshold is established empirically.

34. The apparatus of Claim 26 wherein said first probability threshold is established in a random manner.

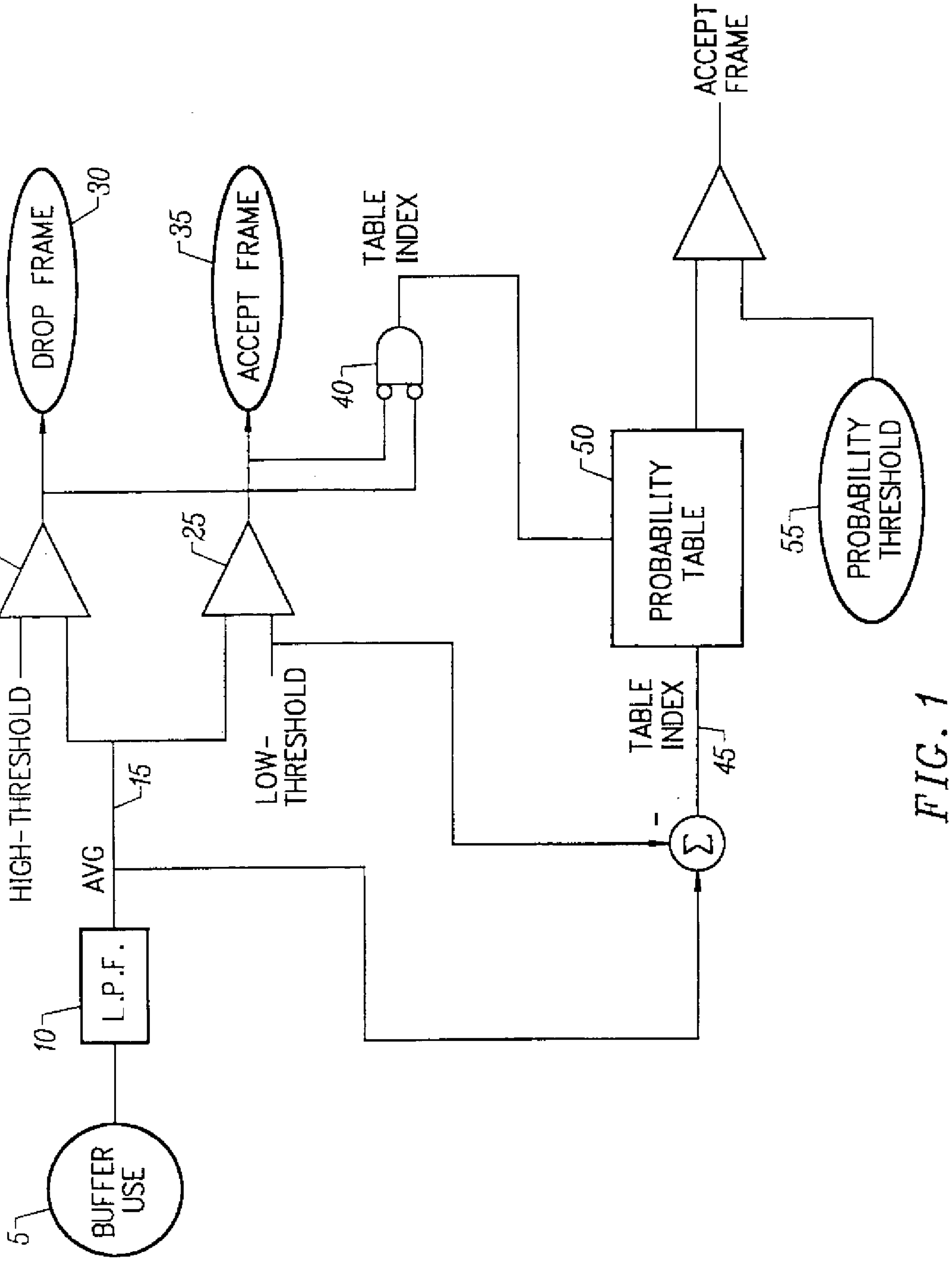

BUFFER USE
5
L.P.F.
10
AVG
15
HIGH-THRESHOLD
LOW-THRESHOLD
20
25
DROP FRAME
30
ACCEPT FRAME
35
40
TABLE INDEX
TABLE INDEX
45
Σ
-
PROBABILITY TABLE
50
PROBABILITY THRESHOLD
55
ACCEPT FRAME

FIG. 1

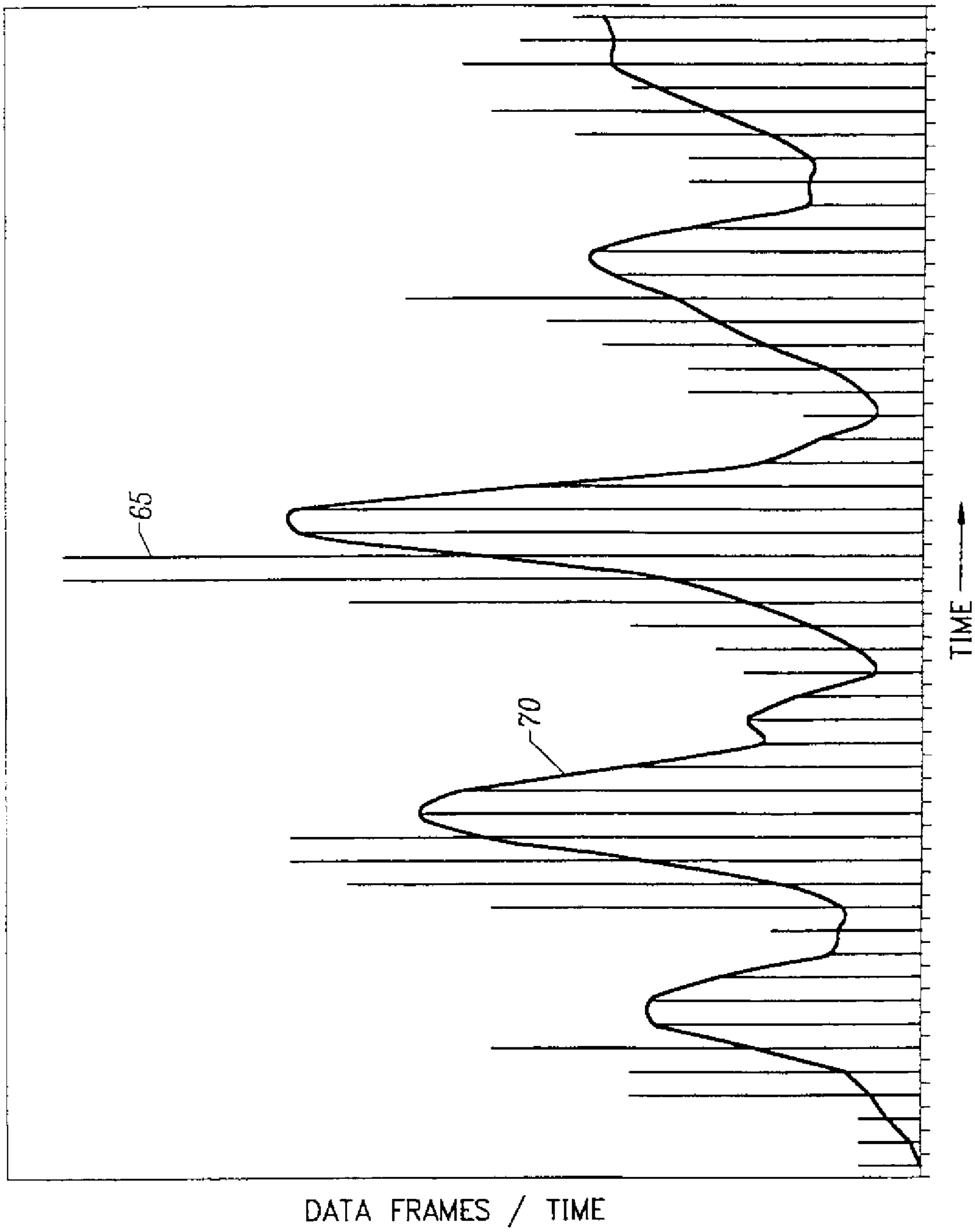


FIG. 2

Sign of Derivative of Buffer Utilization	Sign of Derivative of AVG	Probability Table	Inferences
+	+	1	Both the number of buffers used and the moving average of buffers used are increasing. High likelihood that accepting the frame will result in conjestion.
+	−	2	Number of buffers in use is increasing but the moving average is decreasing.
−	+	3	Number of buffers in use is decreasing but the moving average is increasing.
−	−	4	Both the number of buffers used and the moving average are decreasing. Low probability of congestion if a frame is accepted.

FIG. 3

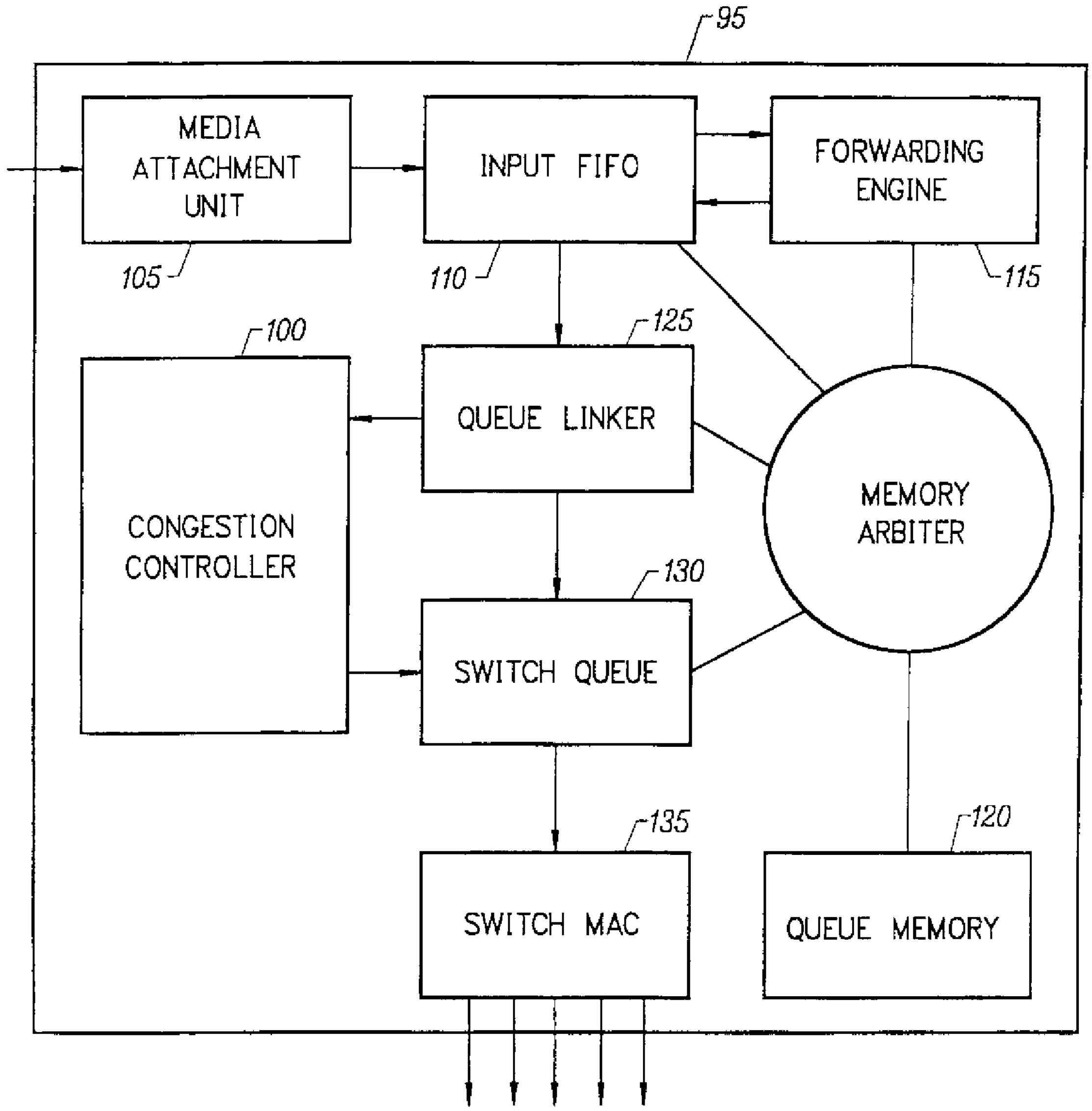

95
MEDIA ATTACHMENT UNIT
INPUT FIFO
FORWARDING ENGINE
105
110
115
100
125
CONGESTION CONTROLLER
QUEUE LINKER
MEMORY ARBITER
130
SWITCH QUEUE
135
120
SWITCH MAC
QUEUE MEMORY

FIG. 4

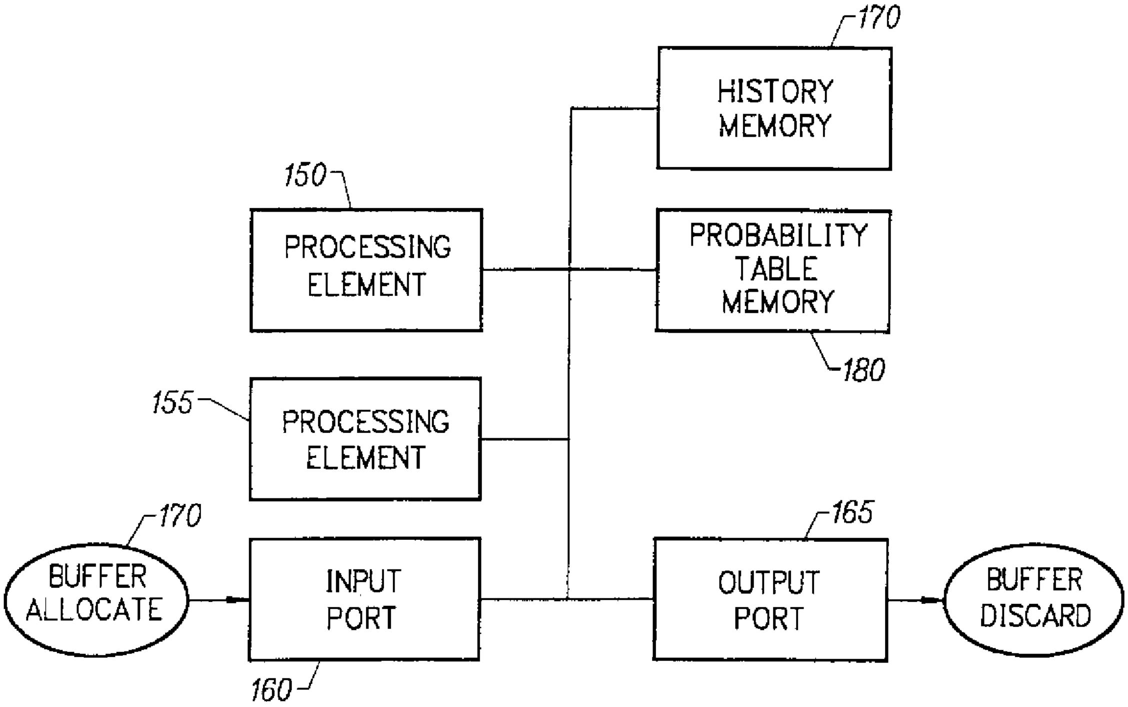

170
HISTORY MEMORY
150
PROCESSING ELEMENT
PROBABILITY TABLE MEMORY
180
155
PROCESSING ELEMENT
170
165
BUFFER ALLOCATE
INPUT PORT
OUTPUT PORT
BUFFER DISCARD
160

FIG. 5

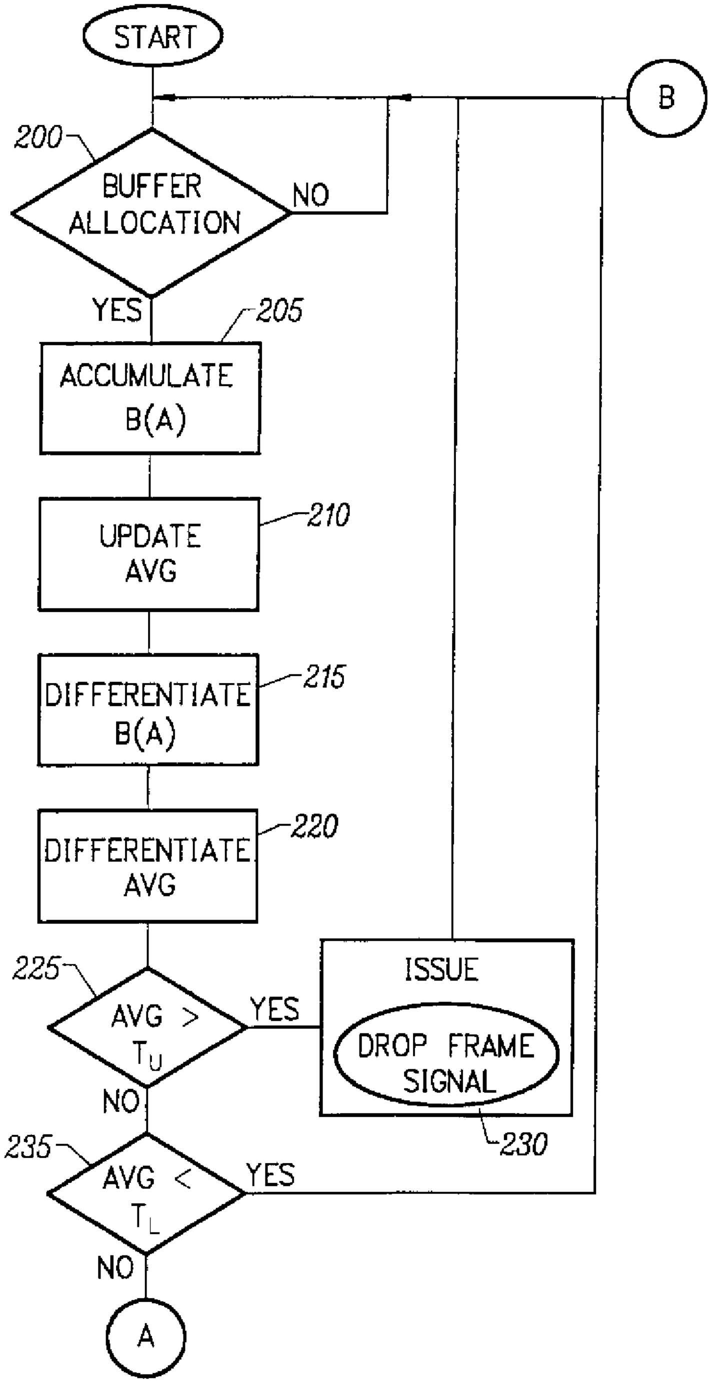

START
B
200
BUFFER ALLOCATION
NO
YES
205
ACCUMULATE B(A)
210
UPDATE AVG
215
DIFFERENTIATE B(A)
220
DIFFERENTIATE AVG
225
AVG > T_U
YES
ISSUE
DROP FRAME SIGNAL
230
NO
235
AVG < T_L
YES
NO
A

FIG. 6A

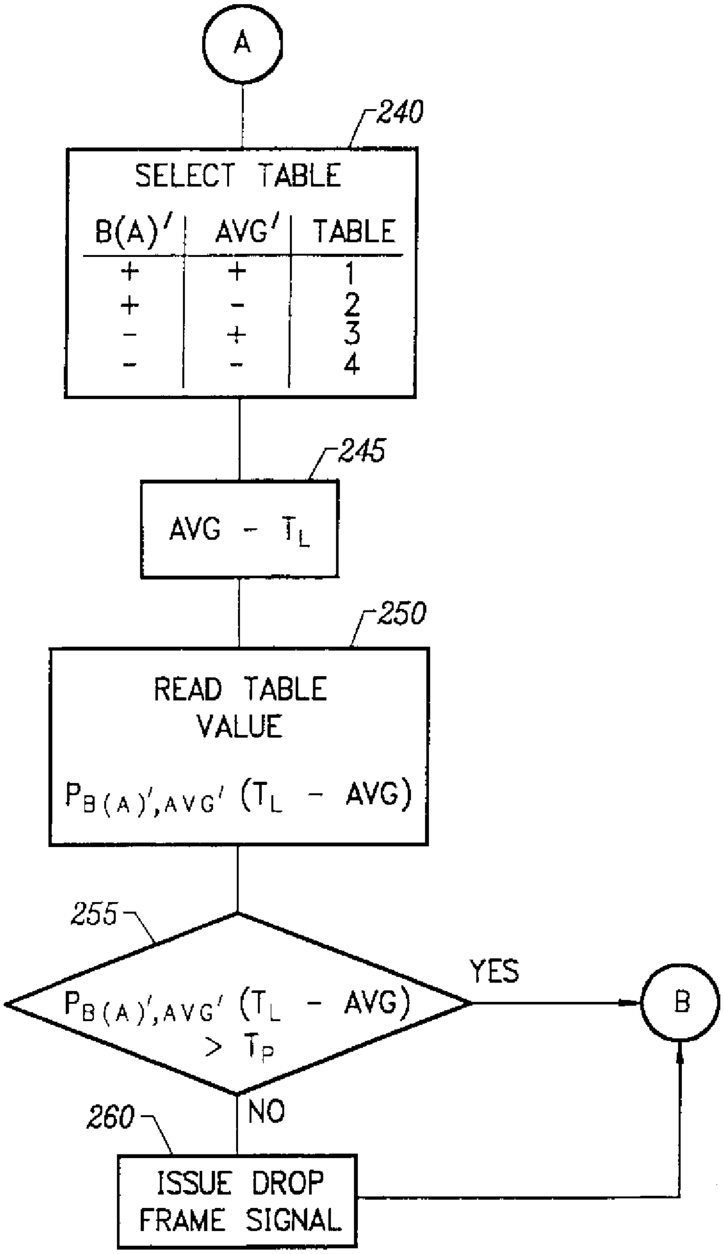

A
240
SELECT TABLE
B(A)' | AVG' | TABLE
+ + 1
+ - 2
- + 3
- - 4
245
AVG - T_L
250
READ TABLE VALUE
$P_{B(A)',AVG'}(T_L - AVG)$
255
$P_{B(A)',AVG'}(T_L - AVG) > T_P$
YES
B
NO
260
ISSUE DROP FRAME SIGNAL

FIG. 6B

INTERNATIONAL SEARCH REPORT

Interna ,l Application No
PCT/US 00/19006

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 H04L12/56

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 H04L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

EPO-Internal, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	FLOYD S ET AL: "RANDOM EARLY DETECTION GATEWAYS FOR CONGESTION AVOIDANCE" IEEE / ACM TRANSACTIONS ON NETWORKING,US,IEEE INC. NEW YORK, vol. 1, no. 4, 1 August 1993 (1993-08-01), pages 397-413, XP000415363 ISSN: 1063-6692 page 400, right-hand column, line 25-43; figure 1 ---	1,9,17, 26
A	US 5 748 901 A (AFEK YEHUDA ET AL) 5 May 1998 (1998-05-05) column 3, line 12-21 column 4, line 53-65 -----	1,9,17, 26

□ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search 22 November 2000	Date of mailing of the international search report 29/11/2000
Name and mailing address of the ISA European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31–70) 340–2040, Tx. 31 651 epo nl, Fax: (+31–70) 340–3016	Authorized officer Dhondt, E

INTERNATIONAL SEARCH REPORT

Information on patent family members

Interna al Application No
PCT/US 00/19006

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5748901 A	05-05-1998	AU 2649197 A	09-12-1997
		WO 9744724 A	27-11-1997